WE CLAIM:

1. A method for identifying a novel nucleic acid molecule encoding a protein of interest comprising:

(i) selecting a specific protein from a first species involved in a regulatory network of interest;

(ii) identifying known proteins that act upstream and downstream in the regulatory network of interest with respect to the specific protein selected;

(iii) constructing the regulatory network of interest from the proteins identified in step (ii);

(iv) for each identified protein, select a domain or motif and search by homology for related proteins in a second species, wherein a related protein is defined as a protein having a homologous domain or motif;

(v) producing a regulatory network for the second species, wherein said regulatory network incorporates the identified related proteins;

(vi) comparing the regulatory network from the first species to the regulatory network of said second species;

(v) identifying a protein present in a regulatory network for one species but absent in the regulatory network of the other species; and

(vi) isolating a nucleic acid molecule encoding the protein identified in step (v) in the species in which it is missing.

2. The method of Claim 1 wherein the nucleic acid molecule encodes human protein.

3. The method of claim 1 wherein the related proteins are orthologs.

4. The method of claim 1 wherein the regulatory pathway is involved in apoptosis.

5. The method of claim 1 wherein the specific protein from the first species is involved in tumor suppression.

6. A method for identifying the affect of a gene knockout on a regulatory pathway comprising the following steps:

(i) identification of the shortest non-oriented pathway connecting two gene products;

(ii) assigning an initial sign value of "-" to the knockout since the knockout gene product is inactive;

(iii) moving along the shortest pathway between the two gene products multiplying the sign with the sign of the next gene product in the pathway, wherein "-" stands for inhibition, "+" stands for induction or activation, and "0" stands for the lack of interaction between two proteins in the specified direction; and

(iv) determining the final sign at the end of the pathway, wherein "-" indicates inhibition and "+" indicates induction or activation of the pathway.

7. A method for identifying a novel nucleic acid molecule encoding a protein of interest comprising:

(i) selecting a gene of interest and searching a database for homologous sequences;

(ii) aligning the homologous sequences identified in step (*i*);

(iii) constructing a gene tree using the sequence alignment;

(iv) constructing a species tree;

(v) imputing the species tree and gene tree into an algorithm which integrates the species tree and the gene tree into a reconciled tree; and

(vi) identifying orthologous genes present in one species but missing in another.

8. The method of claim 7 wherein the following algorithm is used to integrate the species tree and the gene tree into a reconciled tree:

(i) computing the similarity $\sigma(S_{gi},S_{sj})$ for each pair of interior nodes from trees T_g and T_s,

(ii) finding the maximum $\sigma(S_{gi},S_{sj})$;

(iii) saving S_{gi} as a new cluster of orthologs, save $\{S_{gi}\}$ - $\{S_{sj}\}$ as a set of species that are likely to have gene of this kind (or lost it in evolution);

(iv) eliminating S_{gi} from T_g; $T_g := T_g \backslash S_{gi}$;

(v) repeating step (ii)-(iv) until T_g is non-empty.

9. A method for identifying a novel gene comprising the following steps:

(i) defining a motif or domain composition of a gene of interest;

(ii) searching for sequences which correspond to nucleotide sequences in an expression sequence tag database or other cDNA databases using a program such as BLAST and retrieving the identified sequences;

(iii) searching additional databases for expressed sequence tags containing the domains and motifs characteristic for the gene of interest with Hidden Markov Model of domains and motifs identified in step (i);

(iv) identifying nucleotide sequences comprising the gene of interest.

10. The method of claim 9 further comprising using each identified expression sequence tag to search sequence databases for

overlapping sequences for the purpose of assembling longer overlapping stretches of DNA.

SUB B5

11. A method for extracting information on interactions between biological entities from natural-language text data, comprising:

(i) parsing the text data to determine the grammatical structure of the text data ;and

(ii) regularizing the parsed text data to form structured word terms.

12. The method according to claim 11, further comprising preprocessing the data prior to parsing, with preprocessing comprising the step of identifying biological entities.

13. The method according to claim 11, further comprising referring to an additional parameter which is indicative of the degree to which subphrase parsing is to be carried out.

14. The method according to claim 11, wherein said parsing step further comprises segmenting the text data by sentences.

15. The method according to claim 11, wherein said parsing step further comprises:

segmenting the text data by sentences; and

segmenting each of the sentences at identified words or phrases.

16. The method according to claim 11, wherein said parsing step further comprises:

segmenting the text data by sentences; and

segmenting each of the sentences at a prefix.

17. The method according to claim 11, wherein said parsing step further comprises skipping undefined words.

18. The method according to claim 11, wherein said parsing step further comprises:

identifying one or more binary actions and their relationships; and

identifying one or more arguments associated with the actions.

19. The method according to claim 11, further comprising performing error recovery when parsing of the text data is unsuccessful.

20. The method according to claim 19, wherein said error recovery step comprises:

segmenting the text data; and

analyzing the segmented text data to achieve at least a partial parsing of the unsuccessfully parsed text data.

21. The method according to claim 11, wherein said tagging step comprises providing the structured data component in a Standard Generalized Markup Language (SGML) compatible format.

22. A computer system for extracting information on biological entities from natural-language text data, comprising:

(i) means for parsing the natural-language text data; and

(ii) means for regularizing the parsed text data to form structured word terms.

23. The system according to claim 22, further comprising means for preprocessing the data prior to parsing, with the preprocessing means comprising identifying biological entities.

24. The system according to claim 22, further comprising means for referring to an additional parameter which is indicative of the degree to which subphrase parsing is to be carried out.

25. The system according to claim 22, wherein said parsing means further comprises means for segmenting the text data by sentences.

26. The system according to claim 22, wherein said parsing means further comprises:

means for segmenting the text data by sentences; and

means for segmenting each of the sentences at identified words or phrases.

27. The system according to claim 22, wherein said parsing means further comprises:

means for segmenting the text data by sentences; and

means for segmenting each of the sentences at a prefix.

28. The system according to claim 22, wherein said parsing means further comprises means for skipping undefined words.

29. The system according to claim 22, wherein said parsing means further comprises:

means for identifying one or more binary actions and their relationships; and

means for identifying one or more arguments associated with the actions.

30. The system according to claim 22, further comprising means for performing error recovery when parsing of the text data is unsuccessful.

31. The system according to claim 22, wherein said error recovery means comprises:

means for segmenting the text data; and

means for analyzing the segmented text data to achieve at least a partial parsing of the unsuccessfully parsed text data.

32. The system according to claim 22, wherein said tagging means comprises means for providing the structured data component in a Standard Generalized Markup Language (SGML) compatible format.